Using Bayesian Networks to Analyze Expression Data[1]

Friedman, Linial, Nachman, Pe'er

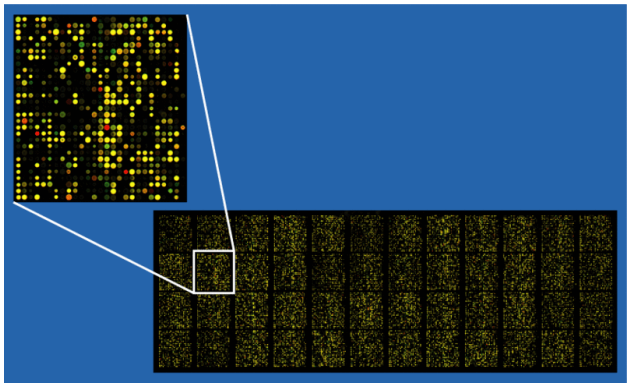
summary by Andreas Kägi <akaegi@student.ethz.ch>

Biological background

In biology proteins play a central role. They are the most important "building parts" of living organisms having important functions (enzymes, structure, regulation, signalling, defence). Molecular biology teaches us that proteins are built in a process called gene expression or protein expression. The information is stored in genes on the DNA, located in the cell nucleus. Since proteins are synthesised in the cytoplasm (liquid outside the nucleus) thee information must be transferred there. This is done by transcribing DNA to mRNA (messenger RNA, a single-stranded helix). The mRNA is transported to the ribosome in the cytoplasm where the protein synthesis take place.

Naturally we want to understand the expression process to better understand organisms in general. More specifically it is interesting to know which factors influence the regulation of gene expression, i.e. the cellular control of the amount and timing of appearance of the functional product (most often a protein) of a gene. Regulation occurs in all phases of gene expression but often in transcription (DNA → mRNA). Thus measuring the concentration of mRNA in a cell gives a rough estimate of the level of gene expression.

mRNA concentration can be measured using the microarray technology. Microarrays contain thousands of cDNA probes (cDNA corresponds to mRNA) that can bind cDNA from a sample.



Link to computer science

Microarray experiments collect enormous amounts of data. Therefore computer aid is necessary to analyse it and find proteins (or other factors) that influence the expression of a certain gene. Problems make the task more challenging: The data is extremely noisy and the data sets are very small. Furthermore mRNA concentrations give only a partial picture of expression activity and data about other key events such as translation (mRNA → protein) is missing.

Previous work was mostly based on clustering algorithms. This has proven to be useful in discovering genes that are co-regulated. This paper has a more sophisticated goal namely discovering the structure of the transcriptional regulation process.

1 http://www.cs.huji.ac.il/labs/compbio/expression

Bayesian Network approach

Friedman et al. do a classical probabilistic approach. They regard a set of random variables and try to obtain the joint probability distribution over it. Achieving that one would be able to ask queries and obtain answers from the joint distribution. The variables mostly correspond to genes and their values to the level of gene expression. Other factors that influence gene expression can be modelled as additional random variables.

Since the model will possibly include thousands of variables, the joint distribution cannot be modelled explicitly. Instead it is modelled using a Bayesian network. These are especially well-suited to this task since one assumes that only a few genes influence the expression of another one. Thus the problem has an inherently local structure which will result in nodes in the Bayesian networks with only a small number of parents.

Local probability model

Using Bayesian networks one has to decide for a local probability model,
i.e. the form of $P(X_i|\boldsymbol{Pa}(\boldsymbol{X}_i))$ (Pa is the parent relation in the network).
Here two models are studied further:

- *Multinomial model* with only discrete variables, where the local distributions are represented as conditional probability tables.
- *Linear Gaussian model* where $P(X_i|\boldsymbol{Pa}(\boldsymbol{X}_i)=\boldsymbol{u}_1..\boldsymbol{u}_k)\sim N(a_0+\sum_i a_i u_i, \sigma^2)$.

Both models have some drawbacks: In the multinomial model you loose information in discretising the expression levels (see also section 4.1 of the paper for a discussion of that). In the Gaussian model only dependencies close to linear can be captured.

Network structure

Learning the network structure, given only observational data is a very hard task, since the number of graphs is super-exponential in the number of variables.
The problem can be formulated as an optimisation problem: Find the Bayesian network B=(G,Θ) that maximises a certain score. The score proposed here is the posterior probability of the graph given the data: $S(G:D)=\log P(G|D)=\log P(D|G)+\log P(G)+C$. We get the first term of the last expression by integrating over all values of theta $P(D|G)=\int P(D|G,\Theta)\,P(\Theta|G)\,d\Theta$. P(G) and P(G|Θ) are just priors.
If one learns from complete data one can *locally compute* this score for each node given only the score values of its parents. (They do not explain this further in this paper but probably in a separate one on this algorithm.[2])

The search space of all DAGs is too big to be searched through with a primitive local search procedure like the greedy hill-climbing algorithm. Therefore Friedman et al. proposes to restrict the number of parents of each node to a set of reasonable candidates. There is a special algorithm to choose this set that is based on the idea that a “good” parent will increase the score contribution of the node (to the total score) a lot. (See section 2.3 for details)

Causal patterns

Once some networks with high score values are selected, one wants to extract causal information from them to get information about the structure of the transcriptional regulation process.
This is done using the concept of Equivalence classes on Bayesian networks:

---

2 *Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate algorithm".* N. Friedman, I. Nachman and D. Pe'er (UAI 99) (can be obtained from site 1)

Two DAGs are equivalent iff they have the same underlying undirected graph and the same v-structure (i.e. converging directed edges into the same node, such as a → b ← c).[3]

Equivalent networks from a class are represented by a PDAG (links present in all members of the class are drawn as a → b, whereas links only present in some networks are drawn as a — b.)

These PDAG class representatives are regarded as "partial" causal networks. Causal networks are identical to Bayesian networks (DAG structure, local probability model) with the difference that a link X → Y is only present if X is a direct cause of Y. (Remember: X → Y and Y → X are equivalent Bayesian networks but not equivalent causal networks!) According to the Causal Markov assumption a causal network can be interpreted as a Bayesian network: The assumption states that a node's value is independent of indirect causes given the direct causes.

The idea is now simple: If all networks in an equivalence class are present, one of them must be the correct causal network. Thus we can safely assume links X → Y in the PDAG representative to be causal links.

Features

Friedman et al. tried to extract two features from the networks, both of which "operate" on pairs of variables. One hopes that these features give insight into the structure of the transcriptional regulation process.

- *Markov relation*: Y is Markov related to X if Y is in the Markov blanket of X. The Markov blanket of X renders X independent of all other variables in the network.
  A Markov relation indicates that two variables are related in some joint biological process.

- *Order relation*: Y is order related to X if X is an ancestor of Y in the PDAG representative (or equivalently in all networks of this class). This indicates that X is a cause of Y which means that X is involved in the regulation of Y. Notice that we cannot be absolutely sure that X is a cause of Y either because the Causal Markov assumption does not hold or because we cannot learn all networks in an equivalence class. Rather we regard it as an indication of a causal relationship.

Verification

Naturally one asks how much confidence one can have in the extracted features. To answer this question they suggested to use the bootstrap method, probably because there is too less data to do crossvalidation. 200 bootstrap data sets G_i are created using resampling with replacement as usual.

Then the confidence in a feature f is calculated as $\mathrm{conf}(f)=\frac{1}{m}\sum_{i=1}^{m} f(G_i)$ where

$f(G_i)=1$ iff feature f can be extracted from data set G_i.

Results

The results section is not that enlightening because the results are not presented that systematically. The data is taken from a survey by Spellman et al. which is presented on their own website very nicely.[4] The data comes from yeast cells. The goal is to identify all genes whose mRNA levels are regulated by the cell cycle. There are 76 measurements of 6177 ORFs (mRNA locations) (i.e. very small data sets!). These measurements were done 6 times with different methods. Spellman himself concluded that 800 genes are regulated during the cell cycle.

To test the robustness of their procedure, Friedman et al. compared their data set to random generated sets. The difference of confidence between the two data sets is very high for the Linear Gaussian model but less significant for the multinomial model. Furthermore, when one compares

3 J. Pearl: Causality, Theorem 1.2.8
4 http://cellcycle-www.stanford.edu

the two methods one sees that the order relation is more stable than the Markov relation. They attributes this to the fact that the Markov relation is a local relation an thus more susceptible to the choice of the local probability model.

From a more biological point of view both relations reveal interesting genes. The order relation reveals dominant genes, i.e. genes that appear before many others in the network. Most of these are biologically interesting.

What they don't do is to compare their results to others with different approaches (i.e. clustering approaches).

Discussion

It seems to make sense to base the analysis of gene expression data on the use of Bayesian and causal networks. What comes to mind is why they have not better made use of the available temporal knowledge from the data. We also learned from Pearl's book that interventions are an important tool when it comes to causal relationships. This issue was not considered here at all.

Other questions are whether changing the local probability model would dramatically change the results and whether the data sets are large enough to trust the results.